Journal of Speech Sciences (9): 93-104. 2020.
Available at: http://revistas.iel.unicamp.br/joss

A METHOD FOR LEXICAL TONE CLASSIFICATION IN AUDIO-VISUAL SPEECH

MENEZES, João Vítor Possamai de[1]*
CANTONI, Maria Mendes[2]
BURNHAM, Denis[3]
BARBOSA, Adriano Vilela[1,4]

[1]Graduate Program in Electrical Engineering, Federal University of Minas Gerais, Brazil
[2]Faculty of Letters, Federal University of Minas Gerais, Brazil
[3]MARCS Institute for Brain, Behavior & Development, Western Sydney University, Australia
[4]Department of Electronic Engineering, Federal University of Minas Gerais, Brazil

---

Abstract: *This work presents a method for lexical tone classification in audio-visual speech. The method is applied to a speech data set consisting of syllables and words produced by a female native speaker of Cantonese. The data were recorded in an audio-visual speech production experiment. The visual component of speech was measured by tracking the positions of active markers placed on the speaker's face, whereas the acoustic component was captured with a high-quality microphone. A pitch tracking algorithm was used to estimate F0 from the acoustic signal. A procedure for head motion compensation was applied to the tracked marker positions in order to separate the head and face motion components. The data were then organized into four signal groups: F0, Face, Head, Face + Head. The signals in each of these groups were parameterized by means of a polynomial approximation and then used to train an LDA (Linear Discriminant Analysis) classifier that maps the input signals into one of the output classes (the lexical tones of the language). One classifier was trained for each signal group. The ability of each signal group to predict the correct lexical tones was assessed by the accuracy of the corresponding LDA classifier. The accuracy of the classifiers was obtained by means of a K-Fold Cross Validation method. The classifiers for all signal groups performed above chance, with F0 achieving the highest accuracy, followed by Face+Head, Face, and Head, respectively. The differences in performance between all signal groups were statistically significant. Our results show that the proposed method is able to assess how well lexical tone can be predicted from different speech modalities. These results are in agreement with previous findings in the literature suggesting that lexical tone can be predicted not only from F0 signals but also, to a lesser degree, from face and head motion signals*.

Keywords: multimodal speech; lexical tone; Cantonese language; statistical learning; Linear Discriminant Analysis.

---

*Corresponding author: joaovmenezes@gmail.com

1 Introduction

Speech is a multimodal phenomenon. While undeniably produced with sounds and gestures, speech is indeed perceived via acoustic and visual information processing. The importance of audiovisual patterns to speech perception came to attention with the works of Sumby and Pollack (1954) and McGurk and MacDonald (1976), and has been well established ever since.

The understanding that speech has a multimodal nature supported the development of new linguistic theories that approached speech production and perception on the basis of articulatory gestures (Brownman and Goldstein, 1986) or body gestures (McNeill, 1981). Speech multimodality also inspired the design of speech technologies able to use non-acoustic speech-related signals to complement acoustic signals, e.g., silent speech interfaces (Denby et al., 2010). However, the development of multimodal speech research brought its own challenges, such as the need for techniques to record and integrate speech signals from different modalities. While the acoustic signal of speech can be easily recorded with any computer or smartphone, the recording of other speech-related signals such as the movement of the vocal tract or the movement of facial articulators require more elaborate and expensive techniques and equipment, e.g., electromagnetic articulography (EMA) for vocal tract, face and head movement and Optotrak for face and head movement (Vatikiotis-Bateson and Ostry, 1995; Yehia, Rubin and Vatikiotis-Bateson, 1998; Tiede et al., 2012).

Another issue is the integration of acoustic and non-acoustic speech signals, because their relation may be direct or indirect. While the movement of the vocal tract affects acoustic speech directly, hand and body gestures do not affect acoustic speech directly, but are part of the multimodal speech being produced (Danner, Barbosa and Goldstein, 2018).

Multimodality can also be addressed in the context of tone languages, which are estimated to comprise 60% to 70% of all languages in the world (Fromkin, 1978) and spoken by more than half of the world's population (Yip, 2002). Tone languages are characterized by the use of lexical tones, which are mainly defined by pitch variations systematically associated with changes in the core meaning of a word (Yip, 2002). Pitch is related to the fundamental frequency (F0) of the acoustic speech signal, which is therefore the primary acoustic parameter used to characterize tones. However, face and head movements were also found to impact lexical tone production and perception. Based only on visual stimuli, both native and non-native speakers of Cantonese differentiated lexical tones with above-chance accuracy (Burnham et al., 2001a; Burnham et al., 2001b). In addition, the perception of Mandarin lexical tones in noisy conditions by both native and non-native perceivers was found to be aided by visual information (Mixdorff et al., 2005; Smith and Burnham, 2012). Relations were also found between different tones and specific head, neck and lip movement patterns of the speaker, but the perceivers could only identify these relations after receiving specific training (Chen and Massaro, 2008). More recently, with the aid of computer vision techniques, it has been suggested that specific lexical tones are related with eyebrow and lip movements (Garg et al., 2019).

In this paper we propose a classification method to assess multimodal speech research questions. The proposed method allows statistical comparisons of acoustic and movement signals and can be used to model the impact of gestures on speech. Here, the method is illustrated by applying it to a multimodal lexical tone classification study, in which tones are identified based not only on the acoustic component of speech, but also on head and face movement.

We argue that the method we propose is capable of assessing the relation between face and head movement and lexical tone. This is done by using the proposed method to quantify how well lexical tones can be determined from head and face movement and then comparing the

results to the case where lexical tones are determined from the F0 signal alone, which is well known to be capable of differentiating lexical tones. The remainder of this paper is organized as follows. Section 2 describes the methodology, covering the data collection procedure, the signal processing techniques applied to the data and the classification methods used. Section 3 describes the results obtained by the application of the proposed method to the experiment's data and the statistical tests applied to them. Section 4 discusses the results. Section 5 summarizes the work and proposes further steps for related research.

2 Methodology

This section describes the data acquisition process, the signal pre-processing procedure and the data analysis.

2.1 Data collection

The data used in this study was collected in an audiovisual speech production experiment conducted in 2013 at the MARCS Institute for Brain, Behavior and Development, Sydney, Australia. The experiment was performed for one female native speaker of Cantonese (age: 20+ years old), a Chinese tone language (Yip, 2002). Acoustic and visual data were recorded while the participant spoke a set of syllables and words that covered the 6 lexical tones of the Cantonese language.

The visual component of speech was measured by tracking 33 active (infrared emitting) markers placed on the speaker's face, head and neck, as shown in Figure 1. The position (x, y, z) of each marker was captured by an NDI Optotrak device (Northern Digital Inc., 2020a) at 60 samples/second, whereas the audio was simultaneously captured by a high-quality microphone and then digitized and recorded at 44100 samples/second by an NDI ODAU[1] device.

The recording was conducted as follows. Initially, the participant had the Optotrak markers attached to her face and then sat in a chair facing the Optotrak device. She was then visually presented with a series of tokens appearing on a screen positioned in front of her. Each token was accompanied by a beep, after which the participant would utter the token shown on the screen. For each beep, a token was recorded. There were 216 different tokens in total, consisting of both monosyllabic words and isolated syllables (which did not constitute a word in Cantonese). Each token was repeated 4 times, resulting in 864 tokens, as shown below:

72 word-tokens, formed by the combination of 12 words (ji, fu [x2], si, sɛ, fɐn, jen, hau, ha:u, jau, soei, wai) with 6 tones;

144 syllable-tokens formed by the combination of 24 syllables with 6 tones. The 24 syllables are formed by the combination of 8 consonants (p^h, p, t^h, t, k^h, k, m, n) with 3 vowels (a, i, u).

These 36 unique tokens (12 words + 24 syllables), which cover most of the consonant-vowel combinations of Cantonese, were selected in order to ensure linguistic generality in the dataset. During the experiment, some of the tokens were not properly recorded and therefore had to be discarded (typically because a marker got detached from the subject's skin or was temporarily occluded or out of the camera's line-of-sight). This resulted in the loss of 30 tokens, reducing the total number of tokens to 834.

[1] The ODAU (Optotrak Data Acquisition Unit) is an optional component of the Optotrak system used to measure analog signals from external sources and convert them to digital format for processing by the main Optotrak unit (Northern Digital Inc., 2020b).

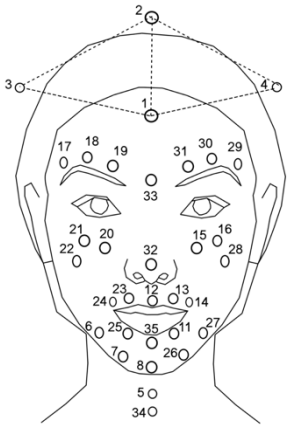

Figure 1: Placement of Optotrak markers on the participant's face.

In Figure 1, Optotrak markers 1 through 4 were attached to a headgear worn by the participant and were meant to capture the rigid body motion of the head. Markers 5 and 34 were placed on the participant's neck. Markers 9 and 10 were inactive during the experiment.

2.2 Pre-processing

The data collected are composed of signal groups in the movement domain and in the audio domain, each one of them receiving a specific processing procedure. Below we describe these procedures.

2.2.1 Audio domain and F0 estimation

The estimation of the F0 curve for each token was obtained using the autocorrelation method (Boersma, 1993) implemented in the Praat software (Boersma and Weenink, 2020). Each recording was individually checked to minimize artifacts in the estimated F0 curve, e.g., pitch halving and discontinuities in the curve. This was done by adjusting input parameters of Praat's F0 estimation algorithm such as *Silence Threshold, Voicing Threshold* and *Octave Cost.*

The F0 estimation procedure is based on a short-term signal processing technique in which the input signal is split into frames of short duration that are then processed individually. The frame size is determined by the minimum F0 value we want to be able to detect, while the time step (frame shift) between adjacent frames is the time distance between F0 estimations. The following values were used for F0 estimation parameters: 1) minimum F0 value of 75 Hz, which is more than enough for a female speaker, resulting in a frame size of 40 ms (Boersma, 1993); and 2) time step of 1/60 s, resulting in one F0 value estimated every 1/60 second, matching the sampling frequency of the movement data (60 samples/second). After this procedure, the acoustic component of speech is no longer represented by the original recorded audio (rate: 44100 samples/second), but rather by the F0 curve (rate: 60 samples/second).

2.2.2 Movement domain and head motion compensation

The measured marker positions consist of two components: one associated with the movement of the face and another associated with the movement of the head. We are interested in the individual contributions of each of these components and, therefore, we need to separate them. This is done through a procedure called head motion compensation, which compensates the measured positions for the head motion. First, the head motion is estimated from markers 1 through 4, whose movement is due only to the movement of the head. After that, the estimated

head movement is removed from the movement of the remaining 29 markers. Figure 2 illustrates the difference between the movement of the markers before and after head motion compensation.

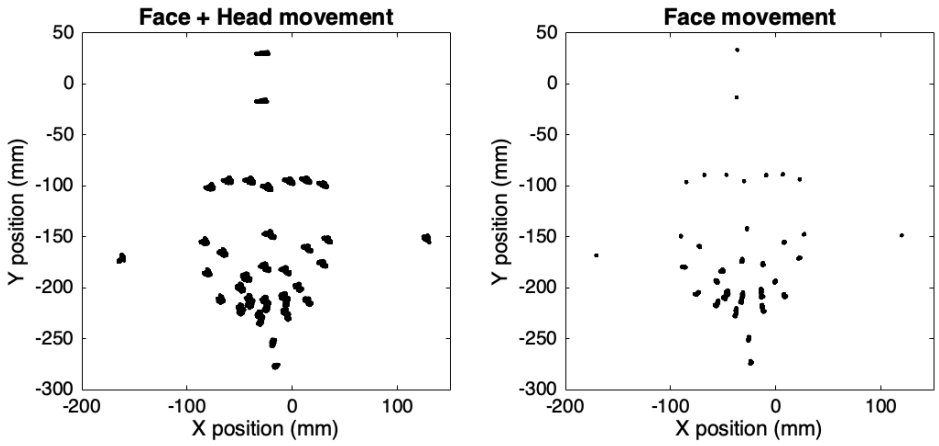

Figure 2: Motion of the markers over 2000 samples (33.33 s). Left: original motion as captured by Optotrak consisting of both face and head components. Right: Face motion component after the head motion compensation procedure. Notice how the head motion is absent on the right pane.

Before the head motion compensation procedure, the visual component of speech was represented solely by the Cartesian coordinates (x, y, z) of the 33 markers as originally measured by Optotrak. After the compensation procedure, the visual component is alternatively represented by two new sets of signals: 1) the head movement and 2) the face movement. Table 1 summarizes the signal groups discussed so far.

Table 1: Summary of the signal groups used as input in the analysis. The F0 signal was obtained with Praat. The Face + Head signals are resulting directly from the measurements of the experiment, whereas Face and Head signals are resulting from the head motion compensation procedure.

Signal group	Composition
F0	1 signal: F_0 value for each frame of the audio signal
Face + Head	99 signals: 33 markers * 3 coordinates (x, y, z)
Face	87 signals: 29 markers * 3 coordinates (x, y, z)
Head	6 signals: 3 translations (x, y, z) + 3 rotations (roll, pitch, yaw)

2.2.3 Polynomial fitting

The individual tokens recorded during the experiment are of different duration. This poses a problem, since the classification method we use (Linear Discriminant Analysis, described in Section 2.3) requires that all input vectors have the same duration. We solve this problem by fitting polynomial curves to each signal of each signal group, thus normalizing signals of different durations to the same duration. More precisely, this procedure allows each signal to be represented by the $p + 1$ coefficients of the polynomial fit to the signal (p is the polynomial order used). Thus, the signals that initially had different durations are now represented by the same number of polynomial coefficients. The change in dimensions is graphically displayed in

Figure 3. The total number of samples for each signal is given by $N = \sum_{i}^{834} n_i$, while the number of signals L for each signal group is given in Table 1.

The polynomial approximation is performed according to the steps below:

1. Before the approximation, the signal is centered at $x = 0$. This ensures that all signals have their middle point at the same *x* coordinate, aiming at the uniformity of the polynomials;
2. Across all signal groups, each signal of each token is approximated individually by a polynomial of order *p* (using the function 'polyfit()' from the 'pracma' package (Borchers, 2019) in R (R Core Team, 2019)).

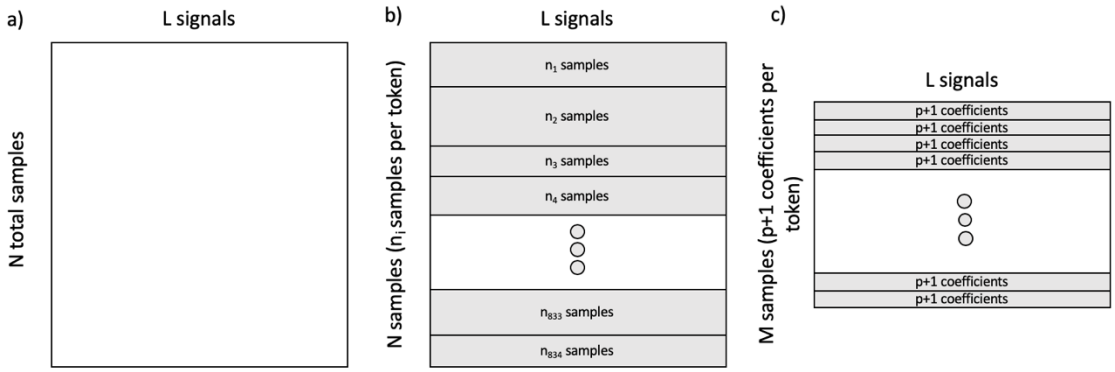

Figure 3: Generic composition of the signals of each signal group described in Table 1. a) The whole data, with N samples of L signals. b) Each token has a different number of samples (n_i) for each signal. c) The data after the fitting of polynomial curves. The total number of samples is now $M = 834 \times (p + 1)$ and the length of each token's signal is (*p+1*).

The procedure of fitting polynomials to signals has a trade-off: while it makes it possible for every token to be described by the same number of samples, it introduces an approximation error, since the original signal cannot usually be represented perfectly by the fit polynomial. Figure 4 shows, for each domain, the mean squared error (MSE) between the original signals and their polynomial approximation as a function of the polynomial order *p*. The MSE between the original signal *x* and its approximation *y* (both with number of samples equal to *n*) is calculated as (James et al., 2013):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$$

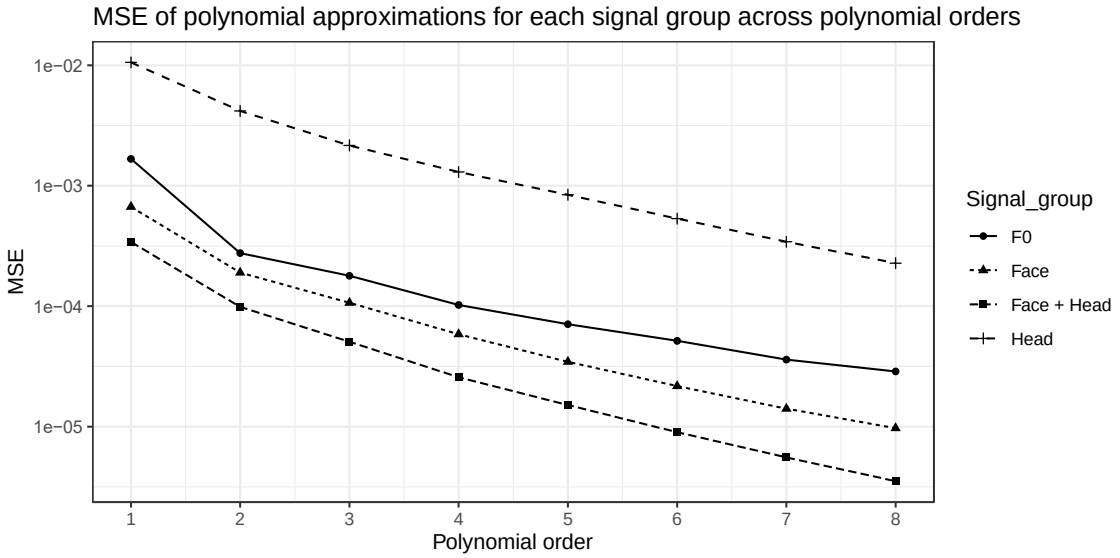

Figure 4: MSE for the polynomial approximations of each signal group. Polynomial orders varied from 1 to 8. Each value is the mean MSE over all signals of each signal group (view Table 1). The y axis is in logarithmic scale for visual clarity.

As the polynomial order increases, more coefficients are used to approximate the signal, which makes the error decrease. We want to use a polynomial order p that results in a low number of coefficients and at the same time introduces a small amount of error. Based on Figure 4, we notice that all error curves are approximately linear from $p = 3$, which means that adding extra coefficients past this point will not decrease the error at the same rate. Thus, we choose $p = 3$ as a good balance between the number of coefficients and the error introduced. In order to maintain uniformity across all domains, all signal groups are approximated by polynomials of the same order ($p = 3$).

2.3 Classification

The perception of lexical tones may be seen as a classification problem: based on an input, which may be auditory, visual or auditory-visual, the perceiver interprets what has been said as one of the possible lexical tones, e.g., 6 different tones in Cantonese. The question is which portions of the input play more important roles in this classification, and with our data, it is possible to measure the classification capacity of 4 different signal groups, as described in Table 1.

A statistical classification method implemented by a computer is not the same as the classification done naturally by humans, who are listening and classifying almost immediately, but the conclusions drawn by it can suggest acoustic and visual features that may play a significant role in lexical tone classification. The next sections describe the classification method used for this study.

2.3.1 Linear Discriminant Analysis

The classification method used in this work is called Linear Discriminant Analysis (LDA), which is a statistical learning method (James et al., 2013). Statistical learning methods work with a set of predictors, here defined as X, and a set of correspondent responses, here defined as Y, trying to estimate a function f that represents the systematic information that is obtained about Y from X, e.g., how the clinical condition of a patient (Y) can be classified based on the

results of certain medical exams (*X*) or, more specifically in our case, how a speech token can be classified into one lexical tone category (*Y*) based on its auditory and/or visual contents (*X*).

The LDA algorithm works in two steps: 1) given the dimensionality of the input predictors *X*, the algorithm finds a smaller set of dimensions that maximizes the separability between classes; 2) it classifies each sample of the data (x_i) into one of the *k* possible classes based on the score of the linear discriminant function, which can be seen as a linear separator between different classes (James et al., 2013).

2.3.2 Training and validation

For training and validation of our LDA model, we performed K-Fold Cross Validation (KFCV), a usual practice in statistical learning methods. Cross validation in general consists of splitting the data in two sets, one for training and one for validation: 1) the training stage comes first and consists of presenting the model with input-output data in order for it to learn the relations between the domains *X* and *Y*. Data used in this stage is from the training set; 2) the validation stage comes after the model has been trained and it checks how good the model is at predicting output from inputs from unseen data. Data used in this stage comes from the validation set. A common metric for the model's performance is classification accuracy, which is the percentage of correctly classified tokens from the validation set.

KFCV is a special type of cross-validation in which the available data is randomly divided in *K* parts, or folds, with approximately the same total number of tokens and balanced number of tokens between all classes. The model is trained with $K - 1$ of these folds and validated with the remaining fold, which was unused for training. This procedure is repeated *K* times so that each fold is used as validation only once. This procedure is commonly used with 5 or 10 folds (James et al., 2013) and can be repeated more than once, each time with a different random division of the folds, in order to assess the variance of the model's behavior.

3 Results

The results described in this section are classification accuracies obtained when the signals of each domain were used as input to an LDA model trained and validated using 60 repetitions of KFCV with $K = 5$ folds. The classifier was also trained and validated with a set composed of one-dimensional uniform random data, so that there is a random accuracy distribution to which the classification accuracies form each domain can be compared. Figure 5 shows the results for each signal group.

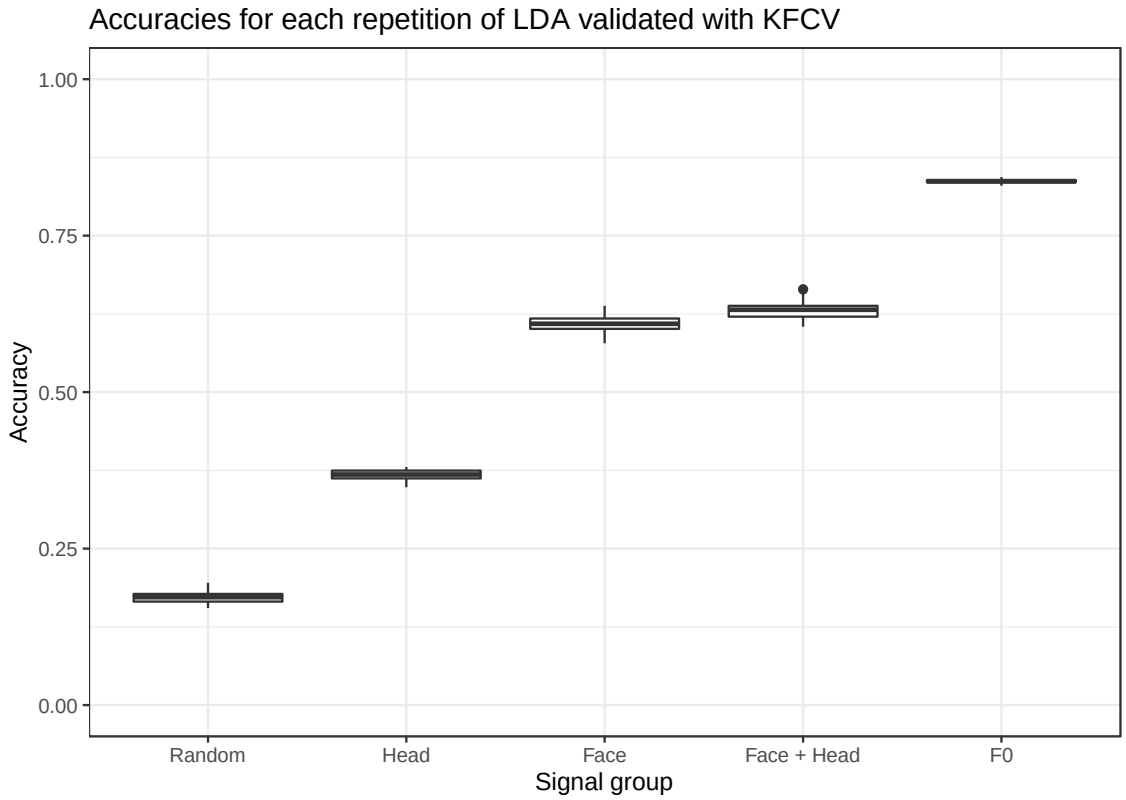

Figure 5: Box plot of the accuracies obtained in LDA classification for each signal group. Since the KFCV was repeated 60 times, each signal group is represented by 60 accuracy values, each of whom is a mean of the $K = 5$ accuracies obtained by each fold.

As expected, the F0 signal presented the highest accuracy values, compared to the movement signals. Visual inspection of Figure 5 seems to indicate that the accuracies attained by the data sets are different. In order to check this, we carried a Kruskal-Wallis test, which confirmed a difference in the means ($p < 0.01$). A non-parametric test was chosen because one of the five sets (Head movement data) did not approach a normal distribution according to a Shapiro-Wilk test ($p < 0.05$) and the homoscedasticity criterium was not met, as indicated by a Bartlett test ($p < 0.05$). With a *post-hoc* analysis, we confirmed the five distributions were significantly different from each other, according to pairwise comparisons with a Dunn test ($p < 0.05$), even the most similar pair of signal groups, Face + Head and Face, whose comparison yielded $p = 0.0177$.

4 Discussion

The accuracy results for all signal groups (F0, Face + Head, Face, Head) were significantly greater than chance, meaning that there is relevant information for lexical tone classification in all signals. To what extent native and non-native Cantonese listeners process this information for lexical tone perception is yet another question. Studies have suggested differences in auditory-visual integration between native and non-native lexical tone perception in Mandarin. For example, in an experiment in which native and tone-naïve non-native speakers were presented with stimuli consisting of congruent and incongruent (e.g., visual information from one lexical tone and auditory information from another tone) combinations of auditory and visual stimuli, Han and colleagues (2020) showed that, while native speakers tend to rely

heavily on the auditory stimulus, non-native speakers integrate auditory and visual information in their processing (although the visual information contributes to perception only marginally and only for specific tones). Other perception studies showed that native and non-native speakers of Cantonese were capable of differentiating between different lexical tones with above chance accuracy when presented with visual-only stimuli (Burnham et al., 2001a; Burnham et al., 2001b). The question of how auditory-visual stimuli influence lexical tone perception is not yet resolved, but our results shed further light on the matter: a statistical approach was able to differentiate between lexical tones with above chance accuracy based only on visual stimulus (face and head movement).

Regarding the relation of the accuracies across different signal groups, F0 classification showed the greatest accuracy, which is expected, since fundamental frequency is the main acoustic correlate of lexical tone. The high accuracy values attained by this signal group should reflect the classifier accuracy itself, meaning that it performs well with appropriate data. Perfect or almost perfect accuracy, which is expected from native speakers under normal conditions, i.e., no noise, was not achieved by our classifier, suggesting that its statistical process for tone classification does not function the same way as human perception. However, the precise differences between the classifier and human perception performance remain to be explicated as it has been found that native tone language and non-native tone language speakers perform similarly and not exactly with 100% accuracy in tone perception (Burnham et al., 2015). Future studies of classifier performance under these conditions will provide further information between classifier and human performance.

On the movement signal groups, all three conditions showed greater accuracy than chance (the random distribution). First, head movement showed significant independent augmentation of tone perception, thus showing that head movement alone contributes to lexical tone classification. While head movement has been found to be related to F0 (Vatikiotis-Bateson and Yehia, 1996; Yehia, Kuratate and Vatikiotis-Bateson, 2002), it has also been reported to be a correlate of longer prosodic structures such as pitch-accent (Krahmer and Swerts, 2007). These results show that head movement is also related to lexical tone classification (and maybe perception). Second, face movement alone contributed to tone classification and did so significantly more than head motion alone. This may be because face motion is a richer, more complex signal than head motion. Lexical tone identification has been related to the intensity of movements such as those from the mouth, chin and neck (Chen and Massaro, 2008) and also of the eyebrows and lips (Garg et al., 2019). Finally, there was a small but significant augmentation of face alone classification by head movement, showing that both face and head movement contribute independently and in an integrated manner to lexical tone classification.

5 Summary

In this work, we proposed a method for lexical tone classification in audio-visual speech. The proposed method resulted in quite high accuracy when predicting lexical tones from the F0 signal, which is the signal most associated with tones. For the other signals (Face + Head, Face, Head), the performance of the method was below F0, but considerably above chance. This shows that there is visual information that contributes to lexical tone classification from both the face and the head movement. The proposed method also might be applied in other tasks, e.g., consonant type or speech act classification, given that the input signals are appropriate for the task, independent of their modality.

Some ramifications that future work can explore are: 1) considering another possible approach for making all tokens the same length, like dynamic time warping (DTW) (Rabiner

and Juang, 1993), to check if approximation errors would be smaller; 2) extending the data to more subjects, as the data we worked with in this study was produced by only one speaker; and 3) extending the data to more languages. In this regard, it has been found that there is augmentation of the acoustic signal by visual speech information not only for native, but also for non-native tone, and even non-tone speakers (Burnham et al., 2015). Further modeling across different languages to incorporate such factors would be informative.

Acknowledgment

Support for this work was provided by CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil) to João Vítor Possamai de Menezes; by FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais, Brazil) to Maria Mendes Cantoni (process APQ-02997-18) and Adriano Vilela Barbosa (process APQ-03701-16); and by Australian Research Council grants to Denis Burnham: 'Making speech three-dimensional: Adding tone to consonant- and vowel-based speech perception and language acquisition research, quantification and theory' (ARC, Discovery DP0988201) and 'Watch my Lips: Perception and Production of Tone and Prosody by Humans and Machines' (ARC, Discovery DP0211947).

REFERENCES

1. Boersma P. *Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Samples Sound*. Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17, 97-100, 1993.

2. Boersma P, Weenink D. *Praat: doing phonetics by computer* [Computer program]. Version 6.1.15, retrieved 20 May 2020 from http://www.praat.org/, 2020.

3. Borchers H W. *pracma: Practical Numerical Math Functions.* R package version 2.2.5. https://CRAN.R-project.org/package=pracma, 2019.

4. Brownman C P, Goldstein L M. *Towards an Articulatory Phonology*. Phonology Yearbook, Vol 3, 219-252, 1986.

5. Burnham D, Ciocca V, Stokes S. *Auditory-Visual Perception of Lexical Tone*. INTERSPEECH, 2001.

6. Burnham D, Lau S, Tam H, Schoknecht C. *Visual Discrimination of Cantonese Tone by Tonal but Non-Cantonese Speakers, and by Non-Tonal Language Speakers.* AVSP 2011 International Conference on Auditory-Visual Speech Processing, 2001.

7. Burnham D, Kasisopa B, Reid A, Luksaneeyanawin S, Lacerda F, Attina V, Xu Rattanasone N, Schwarz I-C, Webster D. *Universality and language-specific experience in the perception of lexical tone and pitch.* Applied Psycholinguistics, 77, 571-591, 2015.

8. Chen T H, Massaro D W. *Seeing pitch: Visual information for lexical tones of Mandarin-Chinese*. The Journal of the Acoustical Society of America, 123, 2356, 2008.

9. Danner S G, Barbosa A V, Goldstein L. *Quantitative analysis of multimodal speech data*. Journal of Phonetics, 71, 268-283, 2018.

10. Denby B, Schultz T, Honda K, Hueber T, Gilbert J M, Brumberg J S. *Silent speech interfaces*. Speech Communication 52, 270-287, 2010.

11. Fromkin V. *Tone: A linguistic survey.* New York: Academic Press, 1978.

12. Garg S, Hamarneh G, Jongman A, Sereno J A, Wang Y. *Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories*. Speech Communication, 113, 47-62, 2019.

13. Han Y, Goudbeek M, Mos M, Swerts M. *Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners.* Language and Speech 1-21, 2020. DOI: https://doi.org/10.1177/0023830919889995.

14. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Springer. 2013.

15. Krahmer E J, Swerts M G J. *The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception.* Journal of Memory and Language, 57(3), 396-414, 2007.

16. McGurk H, MacDonald J. *Hearing lips and seeing voices*. Nature, 264(12), 746-748, 1976.

17. McNeill D. *Action, thought and language*. Cognition, 10, 201-208, 1981.

18. Mixdorff H, Hu Y, Burnham D. *Visual Cues in Mandarin Tone Perception.* INTERSPEECH, 2005.

19. Northern Digital Inc. *Measurement Sciences*. Optotrak Certus. URL https://www.ndigital.com/msci/products/optotrak-certus/, 2020a.

20. Northern Digital Inc. *Measurement Sciences.* Optotrak Acessories. URL https://www.ndigital.com/msci/products/optical-accessories/, 2020b.

21. R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2019.

22. Rabiner L, Juang H. *Fundamentals of Speech Recognition.* PTR Prentice Hall, 1993.

23. Smith D, Burnham D. *Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants*. The Journal of the Acoustical Society of America, 131, 1480, 2012.

24. Sumby W H, Pollack I. *Visual Contribution to Speech Intelligibility in Noise*. The Journal of the Acoustical Society of America, 26(2), 212-215, 1954. DOI: https://doi.org/10.1121/1.1907309.

25. Tiede M, Bundgaard-Nielsen R, Kross C, Gibert G, Attina V, Kasisopa B, Vatikiotis-Bateson E, Best C. *Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously*. The Journal of the Acoustical Society of America. 128(4), 2459, 2010. DOI: https://asa.scitation.org/doi/full/10.1121/1.3508805.

26. Vatikiotis-Bateson E, Ostry D J. *An analysis of the dimensionality of jaw motion in speech.* Journal of Phonetics, 23, 101-117, 1995.

27. Vatikiotis-Bateson E, Yehia H. *Physiological Modeling of Facial Motion During Speech*. Tech. *Rep. ASJ H-96*, 65, 1-8, 1996.

28. Yehia H, Rubin P, Vatikiotis-Bateson E. *Quantitative association of vocal-tract and facial behavior.* Speech Communication 26, 23-43, 1998.

29. Yehia H, Kuratate T, Vatikiotis-Bateson E. *Linking facial animation, head motion and speech acoustics*. Journal of Phonetics, 30, 555-568, 2002.

30. Yip M. *Tone*. Cambridge University Press, 2002.